FEBS Letters 587 (2013) 3500–3507

FEBS Letters

journal homepage: www.FEBSLetters.org

The role of mRNA-based duplication in the evolution of the primate genome

Qu Zhang

Department of Human Evolutionary Biology, Graduate School of Art and Science, Harvard University, Cambridge, MA, USA

ARTICLE INFO

Article history:
Received 7 June 2013
Revised 24 August 2013
Accepted 30 August 2013
Available online 10 September 2013

Edited by Takashi Gojobori

Keywords:
Primate
Retroposition
Positive selection
Comparative genomic
Intact retrocopy

ABSTRACT

Analysis of the human genome suggests novel genes created by retroposition may play an important role in primate evolution. However, data from non-human primates is still scarce. A comprehensive comparison was thus performed among four primate genomes (human, chimpanzee, orangutan, and macaque), which detects elevated rates of retroposition in both the common ancestor of hominoids and macaques. Further analysis shows approximately 10% of intact retrocopies may be under positive selection and at least 4% of retrocopies become functional copies eventually. Moreover, human intact retrocopies were found enriched in transcription-related functions. Collectively, these findings emphasize the important contribution of retroposition to primate genome evolution.

1. Introduction

Recent studies have illustrated that the emergence of new genes is fundamental to the evolution of lineage-specific traits [1–4], and several molecular mechanisms regarding the origin of new genes have been proposed [5]. As first proposed by Ohno in 1970s [6], new functions could be derived through DNA-based gene duplications. which has been extensively studied as a major source of new genetic material [7]. A second form of new gene formation is mRNA-based retroposition [8], where new intronless retrocopies are generated by reverse transcription of mRNA and integrated into random genomic regions [5,9]. These newly formed retrocopies are normally non-functional (retropseudogenes) since they lack regulatory elements. However, a growing number of studies have demonstrated that retrocopies can evolve into functional genes (retrogenes) by acquiring regulatory elements through various ways [10–13] and have been shown to perform an important role in a range of species, including humans [14–18], mice [14], dogs [19,20], mosquitoes [21], and fruit flies [12,13,22,23].

The increasing number of complete genomes provides an unprecedented opportunity to study retropositions at a large scale. Whole genome-scale studies have shown that a high rate of retroposition during primate evolution [15,24,25], which may have been influenced by L1 retrotransposable elements [26]. Other studies have also shown that there is an excess of X-linked parental genes (genes which give rise to retrocopies) in both mammals and flies [13,14,22], a possible result of meiotic X chromosome inactivation (XCI) [27]. However, it is still at the early stage to compare the evolutionary pattern of retrocopies among primates at a genome-wide scale [16,28], therefore a systematic comparative study is desirable to leverage our knowledge on primate genome evolution, lineage-specific adaptations, and fine time scales of genomic changes [29,30]. Thus, I exploited four fully sequenced primate genomes, human (*Homo sapiens*) [31], chimpanzee (*Pan troglodytes*) [32], orangutan (*Pongo pygmaeus*) [33], and macaque (*Macaca mulatta*) [34]) to study the evolutionary pattern of retropositions in primates.

2. Materials and methods

2.1. Retrocopy screen

Protein and genome sequences for human [31], chimpanzee [32], orangutan [33], and macaque [34] were downloaded from the Ensembl release 49 [35]. To identify retrocopies, a bioinformatic pipeline was performed (Fig. 1). First, Ensembl proteomes were compared against themselves by BLASTP [36] and potential parent-retrocopy pairs were defined when (i) one peptide (the parental gene) has multiple exons, while the other (the retrocopy) has only one exon; (ii) the aligned region covers >70%

E-mail address: quzhang@post.harvard.edu

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2013.08.042

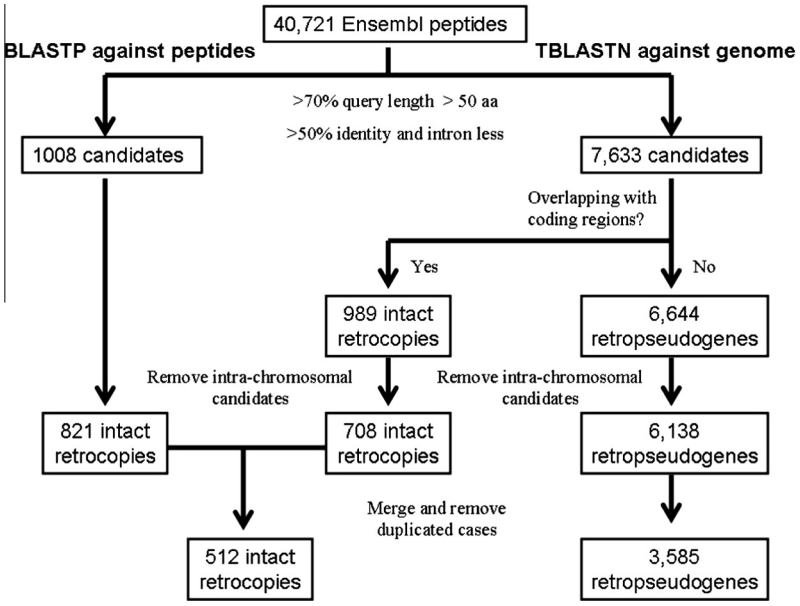

Fig. 1. The flowchart of the pipeline to identify intact retrocopies and retropseudogenes. **N**umbers showed in the figure are for the human genome.

of the parental gene and contains >50 amino acids; and (iii) the alignment has at least 50% similarity. Next, a modified tBLASTn-based approach [15] was applied. Briefly, peptide sequences were aligned against the corresponding genome using tBLASTn, and only matches with more than 50 amino acids in length, >50% sequence identity, and >70% aligned length of the query were retained. The parameters were chosen to keep consistent with previous studies [14,15] to enable later comparisons. If the aligned proportion of the query peptide contains at least two exons, the corresponding genomic hit is denoted as a potential retrocopy. Finally, results from both steps were merged and intra-chromosomal retroposition events were excluded to avoid possible false positives raised by tandem duplications. When overlapping retrocopies were found, only the longer was kept. If there were multiple potential parental genes for a retrocopy, FASTA [37] was used to align them with the retrocopy and the one with the highest alignment was retained. To denote intact retrocopies which may be functional retrogenes, an intact open read frame (ORF) was required with intact start codon and stop codon, as well as absence of frameshift mutation or premature stop codon. Furthermore, intact retrocopies must overlap with coding regions of Ensembl annotated genes which are supported by expressed sequence tags (www.ensembl.org). Retrocopies failed these criteria were denoted as retropseudogenes.

2.2. Ortholog search

For each retrocopy, homologs were searched in two ways. First, I identified flanking genes for each retrocopy (two on each side) and examined whether the orthologs of these flanking genes were also in synteny in other genomes. If these flanking genes were in synteny, then the genomic region between the flanking gene homologs was considered to be the orthologous region; BLASTn was then performed against the homologous region using the retrocopy as query with a cutoff of 10^{-4}. A hit with >70% length and >80% sequence identity with the query sequence was considered to be the ortholog. For retrocopies whose homologs failed to be identified by the previous way, the reciprocal best hit algorithm was used to determine the ortholog.

2.3. Test for accelerated rate of retroposition

A Poisson distribution is used to model the number of identified retropositions in each age class (see results). A likelihood ratio test was constructed to find whether the rate of retroposition on a given branch is significantly different from others. The null hypothesis assumes a constant rate (λ_0, number of retrocopies per million years) for all branches, while the alternative hypothesis assumes one rate (λ_1, number of retrocopies per million years) for a specified branch and a background rate (λ_0) for all other branches. To evaluate whether the Poisson model in the alternative scenario is plausible, index of dispersion (the variance to mean ratio) was calculated for the background branches in the alternative model and all branches in the null model in humans and chimpanzees respectively, and the values in alternative models (0.43 and 1.22 for human and chimpanzee) are much closer to 1 compared to these in null models (9.93 and 9.37), suggesting the Poisson model is a good fit in the alternative model. Therefore the likelihood can be calculated as follows:

$$L_{\text{null}} = \Phi_{i=1}^{n} \frac{e^{-\lambda_0 T}(l_0 T_i)^{K_i}}{K_i!} \quad (1)$$

$$L_{\text{alter}} = \left(\Phi_{i=1}^{n^1 j} \frac{e^{-\lambda_0 T}(\lambda_0 T_i)^{K_i}}{K_i!}\right) \frac{e^{-\lambda_1 T}(\lambda_1 T_j)^{K_j}}{K_j!} \quad (2)$$

where T_i is the length (in million years) for branch i, K_i is the number of retrocopies generated on branch i. Thus twice the difference between log (L_{null}) and log (L_{alter}) is χ^2 distributed with one degree of freedom, and P-value can then be calculated.

2.4. PAML analysis

Pairwise d_S and d_N/d_S statistics for all identified retrocopies and their parental genes were estimated using the YN00 program in PAML [38]. To estimate d_S and d_N/d_S on each branch, a parent-retrocopy pair was aligned with the mouse homolog of the parent gene, and the "free-ratio" model in the codeml program of PAML was used. To test whether retrocopies evolve differently from their parental genes, the "two-ratio" model in codeml with retrocopy set

as the foreground branch was compared with the "one-ratio" model. A branch-site model was used to test the positive selection on retrocopies: M2a with a fixed $\omega_2 = 1$ was compared to M2a with estimated ω_2, as suggested by the PAML manual. The *P*-value for each test was adjusted by Bonferroni multiple test correction.

2.5. PANTHER analysis

Information of functional classification of human protein-coding genes was downloaded from the Protein ANalysis THrough Evolutionary Relationships Classification System (PANTHER) database (http://www.pantherdb.org/). Overrepresentation of a gene in a category was tested by a Fisher's exact test for each category.

2.6. Others

All statistical tests were performed using R (http://www.R-project.org). Plots were made by ggplot2 [39].

3. Results

3.1. Dataset overview

In this study, I applied a comprehensive computational approach (Fig. 1) to search for retrocopies in four primate genomes, and have identified 4097, 3452, 2994 and 3778 retrocopies in humans, chimpanzees, orangutans and macaques, respectively (Table 1 and Supplementary file 1). Among them, 512, 488, 413 and 1177 have intact open reading frame (ORF) and are overlapped with annotated genes. The number of human retrocopies is similar to other studies [15,40]. Distributed on all chromosomes, retrocopies overlapped approximately 2% of annotated genes in primates (Table 1). Notably, the macaque genome has a higher ratio of intact retrocopies compared to other species. However, the d_N / d_S ratio showed the excess of intact retrocopies in macaque may simply reflect mis-classification of certain retropseudogenes as intact retrocopies (Fig. 2). Therefore, macaque is excluded from analyses regarding intact retrocopies.

3.2. Non-constant retroposition rate

Based on an accumulated synonymous substitution $d_S = 0.1$, several previous studies reported a burst of primate retroposition around 38–50 million years ago (MYA), which spans the ancestral catarrhini branch and the ancestral haplorrhini branch [15,24,25]. However, the assumption of a constant molecular clock is often violated by heterogeneity in mutation rates or gene conversions [41], and may lead to less accurate dating. Therefore I estimated the age of each primate retrocopy by directly examining its presence or absence on the phylogenetic tree. To assess the accuracy of this method, a pilot test was performed on 28 retrocopies (Supplementary file 2) previously dated by PCR method [15], and 24 of 28 (86%) genes were assigned to the correct phylogenetic lineage, implying that this method is robust. By this method, retrocopies were divided into several age groups (Table 2): lineage-specific (LS), shared by humans and chimps (HC), shared by humans, chimps and orangutans (HCO), and shared by humans, chimps, orangutans and macaques (HCOM). Retrocopies with ambiguous phylogenetic information or present in both primates and rodents were excluded, leading to 2950, 2508, 2055 and 2480 retrocopies for human, chimpanzee, orangutan and macaque in this analysis (Supplementary file 3). Since the exact calibration points in primate evolution is still under discussion, I used the time estimated by www.timetree.org [42], which is the median value of multiple studies that minimizes random variation in single studies or single methods (Supplementary file 4), and is considered to be close to the true time point. Using divergence time as 6.2 MYA for humans and chimpanzees, 15.3 MYA for African apes and orangutans, 26.8 MYA for apes and the Old World monkeys, and 94.5 MYA for primates and rodents, I found that the rates of retroposition for each group are generally homogenous among species, and the human genome always has the highest rate, which may be attribute to the largest number of human retrocopies due to a high-quality genome. The result clearly demonstrated that along primate evolution, the rate of retroposition is not constant, with an elevation in the common ancestor of great apes (*P*-value $<5 \times 10^{-30}$ in humans, chimpanzees, and orangutans, likelihood ratio test, d.f. = 1). It is also observed that a recent slowdown of retroposition events in chimpanzees and orangutans. In contrast, macaques show a second elevation of retroposition rate after the split of the Old World monkeys and apes (*P*-value = 4.7×10^{-76}, likelihood ratio test, d.f. = 1).

Table 1
Overview of retrocopies in four primate genomes.

Species	# Retrocopy	# Intact retrocopy (%)	% Gene[a]
Human	4097	512 (12.5)	2.3%
Chimpanzee	3452	488 (14.1)	2.5%
Orangutan	2994	413 (13.8)	2.2%
Macaque	3778	1177 (31.2)	5.6%

[a] The percentage of annotated genes that overlap with intact retropcies.

3.3. Evolutionary analysis of retrocopies

The ratio of the non-synonymous substitution rate to the synonymous substitution rate (ω) is a widely used index for understanding evolutionary history of a gene. To study the evolutionary pattern of retrocopies, pairwise ω was estimated for each retrocopy-parental gene pair by YN00 program in PAML, and different patterns were found for intact retrocopies and retropseudogenes (Fig. 2). The distribution of ω for intact retrocopies centered at a peak less than 0.5 for great apes, while the mean ω for retropseudogene pairs is between 0.5 and 1, which is in agreement with the hypothesis that a majority of intact retrocopies are functional and subject to selective constraints, and retropseudogenes are evolving neutrally. Moreover, the clear separation between these two sets also suggests that the intact retrocopy set is enriched for true retrogenes.

Since retrocopies with functional relevance may subject to positive selection [15,43], branch specific ω for each retrocopy and its corresponding parental gene was estimated by adding mouse homologs as the outgroup. Due to the difficulty in identifying mouse homologs for some retrocopies, only a subset of retrocopies was investigated here (Supplementary file 5). First, I tested whether retrocopies evolve differently from their parental genes by comparing the "one-ratio" model with the "two-ratio" model in codeml. The result showed that approximately 20% intact retrocopies were evolving significantly faster than their parents after multiple test correction, in comparison with ~27% retropseudogenes (Table 3). Since both the positive selection and the relaxation of functional constraint could lead to the observed rapid evolution in retrocopies, I next utilized the branch-site model test in codeml to detect positive selection in retrocopies. This test compares the alternative hypothesis that some sites are positively selected and have ω value larger than 1 against the null hypothesis that those sites have ω equal to 1. Approximately 25% intact retrocopies were found under positive selection after Bonferroni correction (Table 4). Since there are only three lineages used in the branch-site model, the test may have limited power and could generate false positive results. However, the proportion of positively selected retropseudogenes could be a naïve estimate of the false discovery rate (FDR)

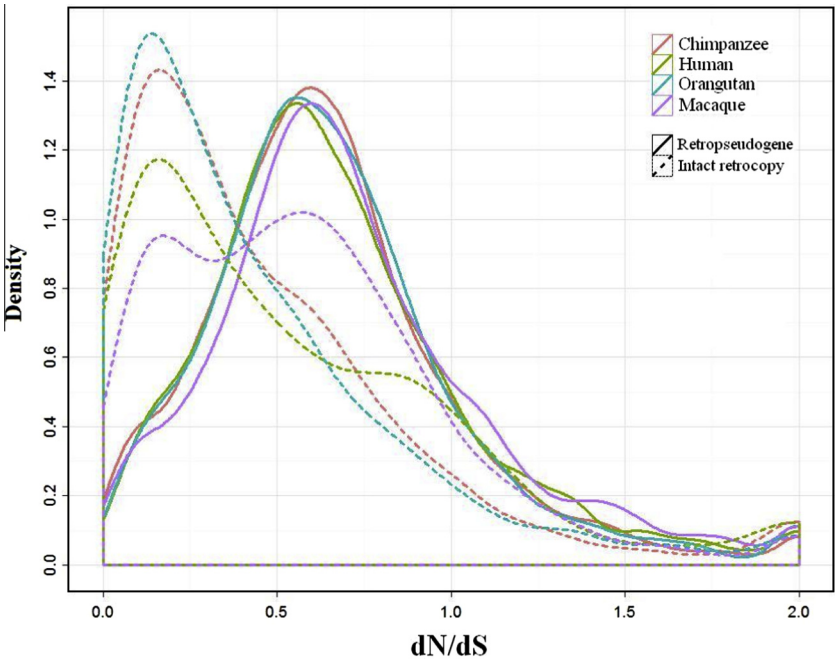

Fig. 2. The distribution of pairwise dN/dS ratio between retrocopies and their parental genes. Dashed lines represent retrogenes, and solid lines represent retropseudogenes. A clearly bimodal pattern is observed for intact retrocopies in macaques, indicating possible contamination of retropseudogenes.

Table 2
Estimated rate of retropositions in four primate genomes.

	HCOM (26.8, 94.5)[a]	HCO (15.3, 26.8)	HC (6.2, 15.3)	LS (0, 6.2)
Human	26.7 (1730)[b]	64.1 (737)	33.6 (306)	28.5 (177)
Chimpanzee	22.5 (1508)	55.5 (638)	27.9 (254)	17.4 (108)
Orangutan	18.3 (1363)	47.0 (540)		9.9 (152)[c]
Macaque	19.8 (1343)			42.4 (1137)[d]

[a] The abbreviation for groups defined in the text. HCOM: retrocopies shared by humans, chimps, orangutans and macaques; HCO: retrocopies shared by humans, chimps and orangutans; HC: retrocopies shared by humans and chimps; LS: lineage-specific retrocopies. Estimated time period for each group is in the unit of million years ago (MYA).
[b] Estimated rate of retropositions as the number of retrocopies per million years (MY).
[c] The time period is [0, 15.3].
[d] The time period is [0, 26.8].

Table 3
Comparison between one-ratio model and two-ratio model.

Species	Intact retrocopy			Retropseudogene		
	$\omega_0 > \omega_1$ (%)[a]	$\omega_0 < \omega_1$ (%)[b]	Total[c]	$\omega_0 > \omega_1$ (%)	$\omega_0 < \omega_1$ (%)	Total
Human	26 (23.6)	1 (0.9)	110	592 (29.2)	8 (0.4)	2028
Chimpanzee	23 (18.4)	2 (1.6)	125	468 (27.7)	7 (0.4)	1691
Orangutan	22 (22.9)	1 (1.0)	96	333 (24.0)	5 (0.4)	1389

[a] Retrocopy-parent pair with a significant LRT *P*-value and has a higher d_N/d_S on the retrocopy branch (ω_0) compared to the parent branch (ω_1).
[b] Retrocopy-parent pair with a significant LRT *P*-value and has a lower d_N/d_S on the retrocopy branch (ω_0) compared to the parent branch (ω_1).
[c] The total number of retrocopy-parent-outgroup trio used.

Table 4
Detection of positive selection using branch-site model in PAML.

Species	Intact retrocopy		Retropseudogene		*P*-value[c]
	Sig (%)[a]	Total[b]	Sig (%)	Total	
Human	26 (23.6)	110	275 (13.6)	2028	<0.01
Chimp	24 (19.2)	125	219 (13.0)	1691	<0.05
Orangutan	29 (30.2)	96	195 (14.0)	1389	<0.001

[a] Retrocopy-parent pair with a significant LRT *P*-value.
[b] The total number of retrocopy-parent-outgroup trio used.
[c] The statistical difference of the proportion of positively selected genes between intact retrocopy and retropseudogenes. *P*-value is calculated by Fisher's Exact Test.

since they are expected to evolve neutrally. Therefore I performed the same analysis on the retropseudogene set and found ~14% retropseudogenes under positive selection, which was utilized as the FDR. After correcting the FDR, the result suggests there may be more than 10% intact retrocopies showing signals of positive selection.

3.4. Rate of retrocopies becoming functional

Newly derived retrocopies generally lack regulatory elements and are considered "dead on arrival" [24,44–46]. But the finding

of retrogenes suggests retrocopies can become functional. Thus one question of specific interest is the proportion of retrocopies that become functional. To obtain a conservative estimate of the ratio of retrogene to retrocopy, potential functional retrogenes were defined as intact retrocopies under selective constraint ($\omega < 0.5$). It should also be notified that the retrocopy screening method is based on sequence similarity, which has limited power to detect old retrocopies, especially non-functional ones. Additionally, sequence saturation could also be a problem in identifying old retrocopies, and a comparison between human and chicken genes showed ~80% of genes are saturated, with a divergence time about 300 million years. Therefore, the ratio was calculated for different d_S values (Fig. 3). As expected, old retrocopies have higher retrogene/retrocopy ratio than young retrocopies, probably due to a slower decay of sequence similarity for functional retrogenes. Retrocopies with $d_S \leqslant 0.6$ (~300 MYA assuming the substitution rate micron for synonymous sites is 10^{-9} per site per year [47]) showed a relatively constant ratio, which indicates few bias in identifying retrogenes and retropseudogenes introduced by sequence saturation or similarity decay. Therefore, estimate based on retrocopies within 300 million years suggest that approximately 4%, or one of every 25 retrocopies eventually became functional.

3.5. Functional analysis of intact retrocopies

Finally, the functional relevance of intact retrocopies was investigated by using the PATHER annotation database. Due to data availability, analyses were only performed for humans. In terms of molecular function, three transcriptional factor-related categories (the KRAB box transcription factor, Zinc finger transcription factor, and Transcription factor, Table 5) are most significantly enriched after Bonferroni multiple test correction. Several other categories (e.g., ribosomal proteins, nucleic acid binding, and histones) are also enriched in intact retrocopies. Since it is possible that the enrichment could simply reflect the enrichment of parental genes, I further studied the functional classification of the parental genes. Of the ten enriched categories, only the three transcriptional factor categories cannot be explained by enriched parental genes (*P*-value $<10^{-6}$, chi-squared test), indicating that intact retrocopies related to transcriptional regulation are preferentially enriched. For biological processes, two transcription-related categories (mRNA transcription regulation and mRNA transcription), as well as two additional categories (cell proliferation and differentiation, and nucleotide and nucleic acid metabolism), were significantly enriched for intact retrocopies after multiple test correction and parental enrichment correction. Although it cannot be directly distinguished whether selective or mutational forces cause the enriched functions, mechanistic changes of gene regulation have been hypothesized to play a central role in mammalian evolution [48,49], and evidence of natural selection in regulatory elements accumulates rapidly [50–54]. Therefore the enrichment of transcriptional regulation in intact retrocopies may be adaptive.

4. Discussion

The functional importance of newly emerged genes have been increasingly recognized in last two decades, as a major contributor to adaptive evolution [10], such as human brain development [55], dog chondrodysplasia [19], mouse amyloid plaque resistance [56], and drosophila male-specific functions [12]. RNA-based retropositions have produced abundant retrocopies across mammalian genomes, hence presenting raw materials for evolution of novelties. In this study, I have performed a comprehensive survey of four primate genomes, and have identified approximately 3000–4000 retrocopies in each species. The number of human retropseudogenes identified here (3585) is smaller than other public databases, such as 8780 in release 68 of pseudogene.org, or 7849 in RCPedia [57] because different parameters were used. For example, method in this study requires > 70% query length, >50% sequence similarity and > 50 amino acids in the aligned region, and inter-chromosomal movement, but the method used in pseudogene.org requires >70% query length coverage and BLAST matches with *E*-values $< 10^{-4}$[24], while the method used in RCPedia has none of the above requirements [18]. Moreover, changes in annotation also account for some differences in the number of retrocopies. Since the underrepresented retropseudogenes in this study may be evolutionary old or unreliable (<50% sequence similarity or <50 amino amids in the alignment), excluding them should have little impact on the analysis of retroposition rate, which only considers primate-specific retrocopies. Additionally, using these pseudogenes may cause uncertainty in sequence alignment and further

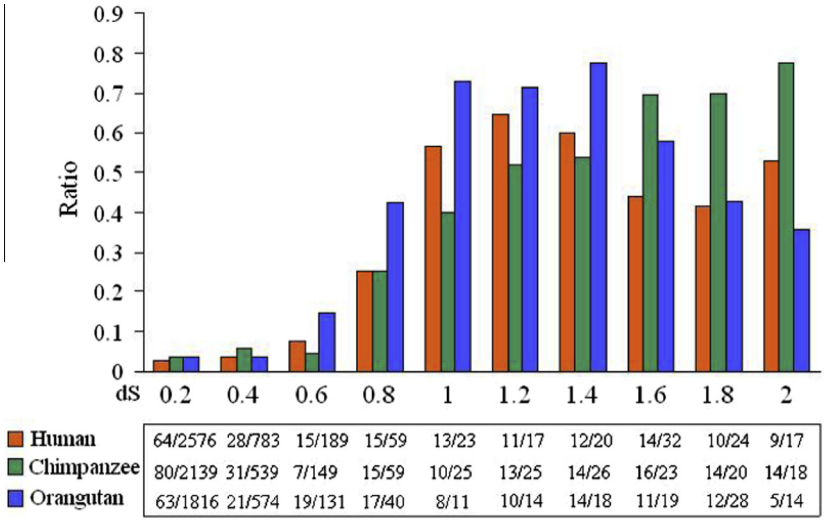

Fig. 3. The proportion of potential retrogenes in retrocopies. The potential retrogenes were defined as intact retrocopies under selective constraint ($\omega < 0.5$). The ratio of retrogenes/retropseudogene was showed across different ages. Ages were divided into bins with $d_S = 0.2$. An increase of ratio is observed in bins with $d_S > 0.6$, which roughly corresponds 300 million years ago (MYA). The table underneath presents the absolute number of retrocopies in each d_S category, in the form of (retrogenes)/(retropseudogenes). It should also be noted that the ratio only represents a minimal estimate due to the stringent parameters used to define potential retrogenes.

Table 5
Functional category enrichment analysis using PANTHER classification.

Category	# Ensembl gene	# Retrogene	*P*-value[a]	# Parent[b]	*P*-value	R/P *P*-value[c]
Molecular function						
KRAB box transcription factor	664	73	<0.001***	6	n.s.	<0.001***
Zinc finger transcription factor	1127	76	<0.001***	8	n.s.	<0.001***
Transcription factor	2892	90	<0.001***	23	n.s.	<0.001***
Ribosomal protein	469	30	<0.001***	71	<0.001***	n.s.
Nucleic acid binding	4478	92	<0.001***	123	<0.001***	n.s.
Histone	512	21	<0.01**	8	n.s.	n.s.
Isomerase	227	13	<0.01**	10	n.s.	n.s.
Small GTPase	317	14	<0.05*	13	<0.05*	n.s.
Other isomerase	106	8	<0.05*	4	n.s.	n.s.
Biological process						
Cell proliferation and differentiation	1461	54	<0.001***	19	n.s.	n.s.
mRNA transcription regulation	2037	65	<0.001***	19	n.s.	<0.01**
Nucleotide and nucleic acid metabolism	5595	123	<0.001***	61	n.s.	n.s.
mRNA transcription	2775	71	<0.001***	26	n.s.	0.05
Protein biosynthesis	625	27	<0.001***	76	<0.001***	n.s.
Nuclear transport	159	12	<0.001***	4	n.s.	n.s.
Protein folding	240	14	<0.001***	14	<0.001***	n.s.
Intracellular protein traffic	1558	36	<0.05*	28	n.s.	n.s.
Chromatin packaging and remodeling	730	21	<0.05*	8	n.s.	n.s.
mRNA processing	415	15	<0.05*	10	n.s.	n.s.

[a] Fisher's exact test is conducted to test the enrichment of a functional category. *P*-value is reported after multiple test correction.
[b] Number of parental genes.
[c] Fisher's exact test is conducted to test the enrichment of a functional category for intact retrocopies by comparing to parental genes. *P*-value is reported after multiple test correction.

evolutionary rate estimate due to their limited sequence similarity and length, thus discarding them in the analyses should give more reliable result than using them. It should also be noted that a filter of selective constraint ($\omega < 0.5$ between the parental genes and intact retrocopies) has been applied in some studies [58]. However, this filter may be too stringent too miss true retrogenes such as *PIPSL*[59], *TMEM183A* or *c1orf37*[43], *GMCL2*, *NACA2* and *PABP3*[15] with $\omega > 1$. Therefore it is not used in this study and three of the five positively selected genes (*PIPSL*, *NACA2* and *PABP3*) were also discovered here. Furthermore, it should be kept in mind that result presented is a minimum estimate of total retrocopies, as retrocopies formed a long time ago may be undetectable due to extreme sequence divergence. Incomplete genome annotations may also reduce the number of retrocopies that can be found.

Results from both this study and previous studies [4,15] indicated that retroposons were more active during some periods during primate evolution. Estimates based on the rate of synonymous change suggested the period is approximately 38–50 MYA [15,25]. Nonetheless, an evolutionary analysis of long interspersed nuclear element 1 families (LINE-1), which are important retroposons in primates, demonstrated that the four most active LINE-1 (L1) families have been extensively amplified between 40 and 12 MYA [60]. Since the heterogeneity in mutation rate could result in biased estimate, a phylogeny-based dating method was applied, along with calibration points estimated from multiple studies and methods. The result showed clear evidence for a burst of retroposition rate in the common ancestor of apes, which is more consistent with the estimated amplification burst from the L1 families [60]. Nevertheless, it could not be completely excluded the impact of uncertainty in calibration point estimates, and a more accurate estimate of speciation time would help confirm the finding here. Additionally, the second burst of retropositions in the macaque genome could be the result of abundant active L1 elements, for instance, ~19,000 L1 elements specific ton the macaque genome [61], though it should also be kept in mind that macaques have shorter generation time compared to apes, the large number of macaque-specific retrocopies may reflect the generation time effect.

It has been long thought that mRNA-based duplicates were just junk DNA and subject to pseudogenization and decay, as they often lack regulatory elements. However, several peculiar properties make a retrocopy, if it becomes active fortuitously, more likely to evolve novel functions compared to DNA-based duplicates [9,62]. First, promoter regions from parental genes are often lost in retrogenes, and the recruitment of new regulatory elements may evolve new function sin retrogenes. Second, retrogenes inserted into existing coding regions change the host gene structure drastically and may result in novel functions. Third, newly derived retrogenes are usually located on different chromosomes from their parents, and may have spatially and temporally distinguished expression pattern. Thus, new genes derived through retroposition provide important source for species-specific adaptations. The findings that approximately 10% of potentially functional retrocopies are under positive selection and their enrichment in transcription-related functions may further highlight the important role of retroposition in primate evolution.

Acknowledgements

I thank Maryellen Ruvolo, Amir Karger, Amanda Lobell, Elizabeth Brown, Justin Gerke, Daniel Green, and Qi Zhou for help on previous versions of this work. I also thank the editor and two anonymous reviewers for helpful comments on an earlier version of this manuscript. This work was supported by the Department of Human Evolutionary Biology, Harvard University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.febslet.2013.08.042.

References

[1] Burki, F. and Kaessmann, H. (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat. Genet. 36, 1061–1063.

[2] Bradley, J., Baltus, A., Skaletsky, H., Royce-Tolland, M., Dewar, K. and Page, D.C. (2004) An X-to-autosome retrogene is required for spermatogenesis in mice. Nat. Genet. 36, 872–876.
[3] Zhang, J., Zhang, Y.P. and Rosenberg, H.F. (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet. 30, 411–415.
[4] Paulding, C.A., Ruvolo, M. and Haber, D.A. (2003) The Tre2 (USP6) oncogene is a hominoid-specific gene. Proc. Natl. Acad. Sci. USA 100, 2507–2511.
[5] Long, M., Betran, E., Thornton, K. and Wang, W. (2003) The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. 4, 865–875.
[6] Ohno, S. (1970) Evolution by Gene Duplication, Springer-Verlag, Berlin.
[7] Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. and Wang, W. (2008) On the origin of new genes in Drosophila. Genome Res. 18, 1446–1455. doi:gr.076588.108, pii:10.1101/gr.076588.108.
[8] Brosius, J. (1991) Retroposons–seeds of evolution. Science 251, 753.
[9] Kaessmann, H., Vinckenbosch, N. and Long, M. (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat. Rev. Genet. 10, 19–31. doi: nrg2487, pii:10.1038/nrg2487.
[10] Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res. 20, 1313–1326. doi: gr.101386.109, pii:10.1101/gr.101386.109.
[11] Fablet, M., Bueno, M., Potrzebowski, L. and Kaessmann, H. (2009) Evolutionary origin and functions of retrogene introns. Mol. Biol. Evol. 26, 2147–2156. doi: msp125, pii:10.1093/molbev/msp125.
[12] Betran, E. and Long, M. (2003) Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. Genetics 164, 977–988.
[13] Bai, Y., Casola, C., Feschotte, C. and Betran, E. (2007) Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biol. 8, R11.
[14] Emerson, J.J., Kaessmann, H., Betran, E. and Long, M. (2004) Extensive gene traffic on the mammalian X chromosome. Science 303, 537–540.
[15] Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A. and Kaessmann, H. (2005) Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3, e357.
[16] Yu, Z., Morais, D., Ivanga, M. and Harrison, P.M. (2007) Analysis of the role of retrotransposition in gene evolution in vertebrates. BMC Bioinform. 8, 308.
[17] Baertsch, R., Diekhans, M., Kent, W.J., Haussler, D. and Brosius, J. (2008) Retrocopy contributions to the evolution of the human genome. BMC Genomics 9, 466. doi: 1471-2164-9-466, pii:10.1186/1471-2164-9-466.
[18] Schrider, D.R., Navarro, F.C., Galante, P.A., Parmigiani, R.B., Camargo, A.A., Hahn, M.W. and de Souza, S.J. (2013) Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet. 9, e1003242. doi: 10.1371/journal.pgen.1003242, pii:PGENETICS-D-12-01322.
[19] Parker, H.G., VonHoldt, B.M., Quignon, P., Margulies, E.H., Shao, S., Mosher, D.S., Spady, T.C., Elkahloun, A., Cargill, M., Jones, P.G., et al. (2009) An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 325, 995–998. doi: 1173275, pii:10.1126/science.1173275.
[20] Kaessmann, H. (2009) Genetics. More than just a copy. *Science* 325, 958–959. doi: 325/5943/958, pii:10.1126/science.1178487.
[21] Toups, M.A. and Hahn, M.W. (2010) Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. Genetics 186, 763–766. doi: genetics.110.118794, pii:10.1534/genetics.110.118794.
[22] Betran, E., Thornton, K. and Long, M. (2002) Retroposed new genes out of the X in Drosophila. Genome Res. 12, 1854–1859.
[23] Schrider, D.R., Stevens, K., Cardeno, C.M., Langley, C.H. and Hahn, M.W. (2011) Genome-wide analysis of retrogene polymorphisms in *Drosophila melanogaster*. Genome Res. 21, 2087–2095. doi: 21/12/2087, pii:10.1101/gr.116434.110.
[24] Zhang, Z., Harrison, P.M., Liu, Y. and Gerstein, M. (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 13, 2541–2558.
[25] Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y. and Okada, N. (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol. 4, R74.
[26] Esnault, C., Maestre, J. and Heidmann, T. (2000) Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. 24, 363–367.
[27] Lyon, M.F. (1961) Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature 190, 372–373.
[28] Pan, D. and Zhang, L. (2009) Burst of young retrogenes and independent retrogene formation in mammals. PLoS ONE 4, e5040. doi: 10.1371/journal.pone.0005040.
[29] Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., et al. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302, 1960–1963. doi: 10.1126/science.1088821, pii:302/5652/1960.
[30] Dorus, S., Vallender, E.J., Evans, P.D., Anderson, J.R., Gilbert, S.L., Mahowald, M., Wyckoff, G.J., Malcom, C.M. and Lahn, B.T. (2004) Accelerated evolution of nervous system genes in the origin of Homo sapiens. Cell 119, 1027–1040. doi: S0092867404011432, pii:10.1016/j.cell.2004.11.040.
[31] Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860–921. doi: 10.1038/35057062.
[32] (2005) The Chimpanzee Sequencing and Analysis Consortium. Nature 437, 69–87, doi: nature04072, pii: 10.1038/nature04072.
[33] Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L.V., Muzny, D.M., Yang, S.P., Wang, Z., Chinwalla, A.T., Minx, P., et al. (2011) Comparative and demographic analysis of orangutan genomes. Nature 469, 529–533. doi: nature04072, pii:10.1038/nature04072.
[34] Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., Venter, J.C., Wilson, R.K., et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316, 222–234. doi: 316/5822/222, pii:10.1126/science.1139247.
[35] Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002) The ensemble genome database project. Nucleic Acids Res. 30, 38–41.
[36] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
[37] Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.
[38] Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.
[39] Wickham, H. (2009) Ggplot2: Elegant Graphics for Data Analysis, Springer, New York.
[40] Chen, M., Zou, M., Fu, B., Li, X., Vibranovski, M.D., Gan, X., Wang, D., Wang, W., Long, M. and He, S. (2011) Evolutionary patterns of RNA-based duplication in non-mammalian chordates. PLoS ONE 6, e21466. doi: 10.1371/journal.pone.0021466, pii:PONE-D-11-01810.
[41] Thomas, J.A., Welch, J.J., Woolfit, M. and Bromham, L. (2006) There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. Proc. Natl. Acad. Sci. USA 103, 7366–7371. doi: 0510251103, pii:10.1073/pnas.0510251103.
[42] Hedges, S.B., Dudley, J. and Kumar, S. (2006) TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22, 2971–2972. doi: btl505, pii:10.1093/bioinformatics/btl505.
[43] Yu, H., Jiang, H., Zhou, Q., Yang, J., Cun, Y., Su, B., Xiao, C. and Wang, W. (2006) Origination and evolution of a human-specific transmembrane protein gene, c1orf37-dup. Hum. Mol. Genet. 15, 1870–1875. doi: ddl109, pii:10.1093/hmg/ddl109.
[44] Petrov, D.A., Lozovskaya, E.R. and Hartl, D.L. (1996) High intrinsic rate of DNA loss in Drosophila. Nature 384, 346–349. doi: 10.1038/384346a0.
[45] Jeffs, P. and Ashburner, M. (1991) Processed pseudogenes in *Drosophila*. Proc. Biol. Sci. 244, 151–159. doi: 10.1098/rspb.1991.0064.
[46] Zhang, Z., Carriero, N. and Gerstein, M. (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. 20, 62–67. Doi: S0168-9525(03)00344-5, pii:10.1016/j.tig.2003.12.005.
[47] Yi, S., Ellsworth, D.L. and Li, W.H. (2002) Slow molecular clocks in Old World monkeys, apes, and humans. Mol. Biol. Evol. 19, 2191–2198.
[48] King, M.C. and Wilson, A.C. (1975) Evolution at two levels in humans and chimpanzees. Science 188, 107–116.
[49] Wilson, A.C., Maxson, L.R. and Sarich, V.M. (1974) Two types of molecular evolution. Evidence from studies of interspecific hybridization. Proc. Natl. Acad. Sci. USA 71, 2843–2847.
[50] Andolfatto, P. (2005) Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437, 1149–1152. Doi: 10.1038/nature04107.
[51] Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.D. and Wray, G.A. (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat. Genet. 39, 1140–1144. doi: 10.1038/ng2104.
[52] Torgerson, D.G., Boyko, A.R., Hernandez, R.D., Indap, A., Hu, X., White, T.J., Sninsky, J.J., Cargill, M., Adams, M.D., Bustamante, C.D., et al. (2009) Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet. 5, e1000592. doi: 10.1371/journal.pgen.1000592.
[53] Gaffney, D.J., Blekhman, R. and Majewski, J. (2008) Selective constraints in experimentally defined primate regulatory regions. PLoS Genet. 4, e1000157. 10.1371/journal.pgen.1000157.
[54] Arbiza, L., Gronau, I., Aksoy, B.A., Hubisz, M.J., Gulko, B., Keinan, A. and Siepel, A. (2013) Genome-wide inference of natural selection on human transcription factor binding sites. Nat. Genet. 45, 723–729. Doi: 10.1038/ng.2658.
[55] Zhang, Y.E., Landback, P., Vibranovski, M.D. and Long, M. (2011) Accelerated recruitment of new brain development genes into the human genome. PLoS Biol. 9, e1001179. doi: 10.1371/journal.pbio.1001179 pii:PBIOLOGY-D-11-01233.
[56] Zhang, Y.W., Liu, S., Zhang, X., Li, W.B., Chen, Y., Huang, X., Sun, L., Luo, W., Netzer, W.J., Threadgill, R., et al. (2009) A functional mouse retroposed gene Rps23r1 reduces Alzheimer's beta-amyloid levels and tau phosphorylation. Neuron 64, 328–340. doi: S0896-6273(09)00674-6 pii:10.1016/j.neuron.2009.08.036.
[57] Navarro, F.C. and Galante, P.A. (2013) RCPedia: a database of retrocopied genes. Bioinformatics 29, 1235–1237. Doi: http://dx.doi.org/10.1093/bioinformatics/btt104.
[58] Potrzebowski, L., Vinckenbosch, N., Marques, A.C., Chalmel, F., Jegou, B. and Kaessmann, H. (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol. 6, e80. doi: 10.1371/journal.pbio.0060080 pii:08-PLBI-RA-0019.
[59] Babushok, D.V., Ohshima, K., Ostertag, E.M., Chen, X., Wang, Y., Mandal, P.K., Okada, N., Abrams, C.S. and Kazazian Jr., H.H. (2007) A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. Genome Res. 17, 1129–1138. Doi: gr.6252107 pii:10.1101/gr.6252107.

[60] Khan, H., Smit, A. and Boissinot, S. (2006) Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16, 78–87. Doi: gr.4001406, pii:10.1101/gr.4001406.

[61] Han, K., Konkel, M.K., Xing, J., Wang, H., Lee, J., Meyer, T.J., Huang, C.T., Sandifer, E., Hebert, K., Barnes, E.W., et al. (2007) Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. Science 316, 238–240. doi: 316/5822/238, pii:10.1126/science.1139462.

[62] Long, M., Deutsch, M., Wang, W., Betran, E., Brunet, F.G. and Zhang, J. (2003) Origin of new genes: evidence from experimental and computational analyses. Genetica 118, 171–182.